BANDWIDTH SELECTION IN KERNEL DENSITY ESTIMATION: ORACLE INEQUALITIES AND ADAPTIVE MINIMAX OPTIMALITY

Alexander Goldenshluger, Oleg Lepski

The Annals of Statistics
2011, Vol. 39, No. 3, 1608–1632
DOI: 10.1214/11-AOS883

BANDWIDTH SELECTION IN KERNEL DENSITY ESTIMATION: ORACLE INEQUALITIES AND ADAPTIVE MINIMAX OPTIMALITY

BY ALEXANDER GOLDENSHLUGER[1] AND OLEG LEPSKI

University of Haifa and Université de Provence

We address the problem of density estimation with $\mathbb{L}_s$-loss by selection of kernel estimators. We develop a selection procedure and derive corresponding $\mathbb{L}_s$-risk oracle inequalities. It is shown that the proposed selection rule leads to the estimator being minimax adaptive over a scale of the anisotropic Nikol'skii classes. The main technical tools used in our derivations are uniform bounds on the $\mathbb{L}_s$-norms of empirical processes developed recently by Goldenshluger and Lepski [*Ann. Probab.* (2011), to appear].

1. Introduction. Let X be a random variable in $\mathbb{R}^d$ having density f with respect to the Lebesgue measure. We want to estimate f on the basis of the i.i.d. sample $\mathcal{X}_n = (X_1, \dots, X_n)$ drawn from f. Any $\mathcal{X}_n$-measurable map $\hat{f} : \mathbb{R}^d \to \mathbb{L}_s(\mathbb{R}^d)$ is understood as an estimator of f, and its accuracy is measured by the $\mathbb{L}_s$-risk:

$$\mathcal{R}_s[\hat{f}, f] := [\mathbb{E}_f \|\hat{f} - f\|_s^q]^{1/q}, \qquad s \in [1, \infty), q \geq 1,$$

where $\mathbb{E}_f$ is the expectation with respect to the probability measure $\mathbb{P}_f$ of the observations $\mathcal{X}_n$. The objective is to develop an estimator of f with small $\mathbb{L}_s$-risk.

Kernel density estimates originate in Rosenblatt (1956) and Parzen (1962); this is one of the most popular techniques for estimating densities [Silverman (1986), Devroye and Györfi (1985)]. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a fixed function such that $\int K(x)\,dx = 1$ (we call such functions *kernels*). Given a *bandwidth vector* $h = (h_1, \dots, h_d)$, $h_i > 0$, the kernel estimator $\hat{f}_h$ of f is defined by

$$\hat{f}_h(t) = \frac{1}{n V_h} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(t - X_i), \tag{1}$$

where $V_h := \prod_{i=1}^d h_i$, u/v for $u, v \in \mathbb{R}^d$ stands for the coordinate-wise division, and $K_h(\cdot) := V_h^{-1} K(\cdot/h)$. It is well known that accuracy properties of $\hat{f}_h$ are determined by the choice of the bandwidth h, and *bandwidth selection* is the central problem in kernel density estimation. There are different approaches to the problem of bandwidth selection.

Received September 2010; revised January 2011.

[1]Supported by ISF Grant 389/07.

MSC2010 subject classifications. 62G05, 62G20.

Key words and phrases. Density estimation, kernel estimators, $\mathbb{L}_s$-risk, oracle inequalities, adaptive estimation, empirical process.

The *minimax approach* is based on the assumption that f belongs to a given class of densities $\mathbb{F}$, and accuracy of $\hat{f}_h$ is measured by its maximal $\mathbb{L}_s$-risk over the class $\mathbb{F}$,

$$\mathcal{R}_s[\hat{f}_h; \mathbb{F}] := \sup_{f \in \mathbb{F}} \mathcal{R}_s[\hat{f}_h; f].$$

Typically $\mathbb{F}$ is a class of smooth functions, for example, the Hölder, Nikol'skii or Besov functional class. Then the bandwidth h is selected so that the maximal risk $\mathcal{R}_s[\hat{f}_h; \mathbb{F}]$ (or a reasonable upper bound on it) is minimized with respect to h. Such a choice leads to a deterministic bandwidth h depending on the sample size n, and on the underlying functional class $\mathbb{F}$. In many cases the resulting kernel estimator constructed in this way is *rate optimal* (or *optimal in order*) over the class $\mathbb{F}$. Minimax kernel density estimation with $\mathbb{L}_s$-risks on $\mathbb{R}^d$ was considered in Bretagnolle and Huber (1979), Ibragimov and Has'minskiĭ (1980), Ibragimov and Khas'minskiĭ (1981), Devroye and Györfi (1985), Hasminskii and Ibragimov (1990), Donoho et al. (1996), Kerkyacharian, Picard and Tribouley (1996), Juditsky and Lambert-Lacroix (2004) and Mason (2009) where further references can be found.

The *oracle approach* considers a set of kernel estimators $\mathcal{F}(\mathcal{H}) = \{\hat{f}_h, h \in \mathcal{H}\}$, and aims at a measurable data-driven choice $\hat{h} \in \mathcal{H}$ such that for every f from a *large* functional class the following $\mathbb{L}_s$*-risk oracle inequality* holds:

$$\mathcal{R}_s[\hat{f}_{\hat{h}}; f] \le C \inf_{h \in \mathcal{H}} \mathcal{R}_s[\hat{f}_h; f] + \delta_n. \tag{2}$$

Here $C > 0$ is a constant independent of f and n, and the remainder δ_n does not depend on f. Oracle inequalities with "small" remainder term δ_n and constant C close to 1 are of prime interest; they are key tools for establishing minimax and adaptive minimax results in estimation problems. To the best of our knowledge, oracle inequalities of the type (2) were established only in the cases $s = 1$ and $s = 2$. Devroye and Lugosi (1996, 1997, 2001) established oracle inequalities for $s = 1$. The case $s = 2$ was studied by Massart [2007, Chapter 7], Samarov and Tsybakov (2007), Rigollet and Tsybakov (2007) and Birgé (2008). The last cited paper contains a detailed discussion of recent developments in this area.

The contribution of this paper is twofold. First, we propose a selection procedure for a set of kernel estimators, and establish for the corresponding $\mathbb{L}_s$-risk, $s \in [1, \infty)$, oracle inequalities of the type (2). Second, we demonstrate that our selection rule leads to a minimax adaptive estimator over a scale of the anisotropic Nikol'skii classes (see Section 3 below for the class definition).

More specifically, let $h^{\min} = (h_1^{\min}, \dots, h_d^{\min})$ and $h^{\max} = (h_1^{\max}, \dots, h_d^{\max})$ be two fixed vectors satisfying $0 < h_i^{\min} \le h_i^{\max} \le 1$, $\forall i$, and let

$$\mathcal{H} := \bigotimes_{i=1}^{d} [h_i^{\min}, h_i^{\max}]. \tag{3}$$

Consider the set of kernel estimators

$$\mathcal{F}(\mathcal{H}) = \{\hat{f}_h, h \in \mathcal{H}\}, \tag{4}$$

where $\hat{f}_h$ is given in (1). We propose a measurable choice $\hat{h} \in \mathcal{H}$ such that the resulting estimator $\hat{f} = \hat{f}_{\hat{h}}$ satisfies the following oracle inequality:

$$\mathcal{R}_s[\hat{f}_{\hat{h}}; f] \le \inf_{h\in\mathcal{H}}\{(1 + 3\|K\|_1)\mathcal{R}_s[\hat{f}_h; f] + C_s(nV_h)^{-\gamma_s}\} + \delta_{n,s}. \tag{5}$$

The constants C_s, γ_s, and the remainder term $\delta_{n,s}$ admit different expressions depending on the value of s.

- If $s \in [1, 2)$, then (5) holds for all densities f with $\gamma_s = 1 - \frac{1}{s}$, C_s depending on the kernel K only, and with

$$\delta_{n,s} = c_1(\ln n)^{c_2} n^{1/s} \exp\{-c_3 n^{2/s-1}\}$$

  for some constants c_i, $i = 1, 2, 3$.
- If $s \in [2, \infty)$, then (5) holds for all densities f uniformly bounded by a constant f_∞ with $\gamma_s = \frac{1}{2}$, C_s depending on K and f_∞ only, and with

$$\delta_{n,s} = c_1(\ln n)^{c_2} n^{1/2} \exp\{-c_3 V_{\max}^{-2/s}\}, \qquad V_{\max} := V_{h^{\max}},$$

  for some constants c_i, $i = 1, 2, 3$. We emphasize that the proposed selection rule is fully data-driven and does not use information on the value of f_∞.

Thus, the oracle inequality (5) holds with exponentially small (in terms of dependence on n) remainder $\delta_{n,s}$ (by choice of $V_{\max}$ in the case $s \in [2, \infty)$). We stress that explicit nonasymptotic expressions for C_s, c_1, c_2 and c_3 are available. It is important to realize that the term $C_s(nV_h)^{-\gamma_s}$ is a tight upper bound on the stochastic error of the kernel estimator $\hat{f}_h$. This fact allows to derive rate optimal estimators that adapt to unknown smoothness of the density f. In particular, in Section 3 we apply our oracle inequalities in order to develop a rate optimal adaptive kernel estimator for the anisotropic Nikol'skii classes. Minimax estimation of densities from such classes was studied in Ibragimov and Khas'minskiĭ (1981), while the problem of adaptive estimation was not considered in the literature.

The paper is structured as follows. In Section 2, we define our selection rule and prove key oracle inequalities. Section 3 discusses adaptive rate optimal estimation of densities for a scale of anisotropic Nikol'skii classes. Proofs of all results are given in Section 4.

2. Selection rule and oracle inequalities. Let $\mathcal{F}(\mathcal{H})$ be the set of kernel density estimators defined in (4). We want to select an estimator from the family $\mathcal{F}(\mathcal{H})$. For this purpose, we need to impose some assumptions and establish notation that will be used in the definition of our selection procedure.

2.1. *Assumptions*. The following assumptions on the kernel K will be used throughout the paper.

(K1) The kernel K satisfies the Lipschitz condition

$$|K(x) - K(y)| \le L_K |x - y| \qquad \forall x, y \in \mathbb{R}^d,$$

where $|\cdot|$ denotes the Euclidean distance. Moreover, K is compactly supported, and, without loss of generality, $\operatorname{supp}(K) \subseteq [-1/2, 1/2]^d$.

(K2) There exists a real number $\mathrm{k}_\infty < \infty$ such that $\|K\|_\infty \le \mathrm{k}_\infty$.

Assumptions (K1) and (K2) are rather standard in kernel density estimation. We note that Assumption (K1) can be weakened in several ways. For example, it suffices to assume that K belongs to the isotropic Hölder ball of functions $\mathbb{H}_d(\alpha, L_K)$ with any $\alpha > 0$ [in Assumption (K1) $\alpha = 1$].

Sometimes we will suppose that $f \in \mathbb{F}$, where

$$\mathbb{F} := \left\{ p : \mathbb{R}^d \to \mathbb{R} : p \ge 0, \int p = 1, \|p\|_\infty \le \mathrm{f}_\infty < \infty \right\},$$

and f_∞ is a fixed constant. Without loss of generality we assume that $\mathrm{f}_\infty \ge 1$.

2.2. *Notation*. For any $U : \mathbb{R}^d \to \mathbb{R}$ and $s \in [1, \infty)$ define

$$\rho_s(U) := \begin{cases} 4n^{1/s-1}\|U\|_s, & s \in [1, 2), \\ n^{-1/2}\|U\|_2, & s = 2, \end{cases}$$

and if $s \in (2, \infty)$, then we set

$$\rho_s(U) := D_s \left\{ n^{-1/2} \left(\int \left[\int U^2(t - x) f(x)\, \mathrm{d}x \right]^{s/2} \mathrm{d}t \right)^{1/s} + 2n^{1/s-1}\|U\|_s \right\},$$

where $D_s := 15s/\ln s$ is the best-known constant in the Rosenthal inequality [Johnson, Schechtman and Zinn (1985)]. Observe that $\rho_s(U)$ depends on f when $s \in (2, \infty)$; hence we will also consider the empirical counterpart of $\rho_s(U)$:

$$\hat{\rho}_s(U) := D_s \left\{ n^{-1/2} \left(\int \left[\frac{1}{n} \sum_{i=1}^n U^2(t - X_i) \right]^{s/2} \mathrm{d}t \right)^{1/s} + 2n^{1/s-1}\|U\|_s \right\}.$$

We put also

$$r_s(U) := \rho_s(U) \vee n^{-1/2}\|U\|_2, \qquad \hat{r}_s(U) := \hat{\rho}_s(U) \vee n^{-1/2}\|U\|_2$$

and

$$g_s(U) := \begin{cases} 32\rho_s(U), & s \in [1, 2), \\ \dfrac{25}{3}\rho_2(U), & s = 2, \\ 32\hat{r}_s(U), & s > 2. \end{cases}$$

Armed with this notation we are ready to describe our selection rule.

2.3. *Selection rule.* The rule is based on auxiliary estimators $\{\hat{f}_{h,\eta}, h, \eta \in \mathcal{H}\}$ that are defined as follows: for every pair $h, \eta \in \mathcal{H}$ we let

$$\hat{f}_{h,\eta}(t) := \frac{1}{n}\sum_{i=1}^{n}[K_h * K_\eta](t - X_i),$$

where "$*$" stands for the convolution on $\mathbb{R}^d$. Define also

$$\begin{aligned} m_s(h,\eta) &:= g_s(K_\eta) + g_s(K_h * K_\eta) \qquad \forall h, \eta \in \mathcal{H}, \\ m_s^*(h) &:= \sup_{\eta\in\mathcal{H}} m_s(\eta, h) \qquad \forall h \in \mathcal{H}. \end{aligned} \tag{6}$$

For every $h \in \mathcal{H}$ let

$$\hat{R}_h := \sup_{\eta\in\mathcal{H}}[\|\hat{f}_{h,\eta} - \hat{f}_\eta\|_s - m_s(h,\eta)]_+ + m_s^*(h). \tag{7}$$

The selected bandwidth $\hat{h}$ and the corresponding kernel density estimator are defined by

$$\hat{h} := \arg\inf_{h\in\mathcal{H}} \hat{R}_h, \qquad \hat{f} = \hat{f}_{\hat{h}}. \tag{8}$$

The selection rule (6)–(8) is a refinement of the one introduced recently in Goldenshluger and Lepski (2008, 2009) for the Gaussian white noise model.

REMARKS. 1. It is easy to check that Assumption (K1) implies that $\hat{R}_h$ and $m_s^*(h)$ are continuous random functions on the compact subset $\mathcal{H} \subset \mathbb{R}^d$. Thus, $\hat{h}$ exists and is measurable [Jennrich (1969)].

2. We call function $m_s(\cdot,\cdot)$ the *majorant*. In fact, if ξ_h and $\xi_{h,\eta}$ denote the stochastic errors of estimators $\hat{f}_h$ and $\hat{f}_{h,\eta}$, respectively, that is, if

$$\begin{aligned} \xi_h(t) &:= \frac{1}{n}\sum_{i=1}^{n}[K_h(t - X_i) - \mathbb{E}_f K_h(t - X)], \\ \xi_{h,\eta}(t) &:= \frac{1}{n}\sum_{i=1}^{n}\{[K_h * K_\eta](t - X_i) - \mathbb{E}_f[K_h * K_\eta](t - X)\}, \end{aligned}$$

then it is seen from the proofs of Theorems 1 and 2 below that $m_s(h,\eta)$ uniformly "majorates" $\|\xi_{h,\eta} - \xi_\eta\|_s$ in the sense that the expectation

$$\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} [\|\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)]_+^q$$

is "small."

3. It is important to realize that the majorant $m_s(h,\eta)$ is explicitly given and does not depend on the density f to be estimated. The majorant is completely determined by kernel K and observations, and thus it is available to the statistician.

2.4. *Oracle inequalities*. Now we are in a position to establish oracle inequalities on the risk of the estimator $\hat{f} = \hat{f}_{\hat{h}}$ given by (8). Put

$$A_{\mathcal{H}} := \prod_{i=1}^{d} [1 \vee \ln(h_i^{\max}/h_i^{\min})], \qquad B_{\mathcal{H}} := [1 \vee \log_2(V_{\max}/V_{\min})],$$

where from now on

$$V_{\min} := \prod_{i=1}^{d} h_i^{\min}, \qquad V_{\max} := \prod_{i=1}^{d} h_i^{\max}.$$

The next two statements, Theorems 1 and 2, provide oracle inequalities on the $\mathbb{L}_s$-risk of $\hat{f}$ in the cases $s \in [1, 2]$ and $s \in (2, \infty)$, respectively.

THEOREM 1. *Let Assumptions* (K1) *and* (K2) *hold.*

(i) *If* $s \in [1, 2)$, *then for all* f *and* $n \geq 4^{2s/(2-s)}$

$$\begin{aligned} \mathcal{R}_s[\hat{f}; f] \leq \inf_{h \in \mathcal{H}} \big[& (1 + 3\|K\|_1) \mathcal{R}_s[\hat{f}_h, f] + C_1 (n V_h)^{1/s-1} \big] \\ & + C_2 A_{\mathcal{H}}^{4/q} n^{1/s} \exp\left\{ -\frac{2n^{2/s-1}}{37q} \right\}. \end{aligned} \tag{9}$$

(ii) *If* $s = 2$ *and* $\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2} \leq 1/8$, *then for all* $f \in \mathbb{F}$

$$\begin{aligned} \mathcal{R}_s[\hat{f}; f] \leq \inf_{h \in \mathcal{H}} \big[& (1 + 3\|K\|_1) \mathcal{R}_s[\hat{f}_h, f] + C_3 (n V_h)^{-1/2} \big] \\ & + C_4 A_{\mathcal{H}}^{4/q} n^{1/2} \exp\left\{ -\frac{1}{16q[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}]} \right\}. \end{aligned} \tag{10}$$

Here C_1 *and* C_3 *are absolute constants, while* C_2 *and* C_4 *depend on* L_K, k_∞, d *and* q *only.*

THEOREM 2. *Let Assumptions* (K1) *and* (K2) *hold,* $s \in (2, \infty)$, *and assume that for some* $C_1 = C_1(K, s, d) > 1$

$$n V_{\min} > C_1, \qquad V_{\max} \geq 1/\sqrt{n}.$$

If $n \geq C_2$ *for some constant* C_2 *depending on* L_K, k_∞, f_∞, d *and* s *only, then* $\forall f \in \mathbb{F}$,

$$\begin{aligned} \mathcal{R}_s[\hat{f}; f] \leq \inf_{h \in \mathcal{H}} \big[& (1 + 3\|K\|_1) \mathcal{R}_s[\hat{f}_h, f] + C_3 \mathrm{f}_\infty^{1/2} (n V_h)^{-1/2} \big] \\ & + C_4 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{1/2} [\exp\{-C_5 b_{n,s}\} + \exp\{-C_6 \mathrm{f}_\infty^{-1} V_{\max}^{-2/s}\}], \end{aligned} \tag{11}$$

where $b_{n,s} := n^{4/s-1}$ *if* $s \in (2, 4)$, *and* $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ *if* $s \geq 4$. *The constants* C_i, $i = 3, \ldots, 6$, *depend on* L_K, k_∞, d, q *and* s *only.*

REMARKS. 1. All constants appearing in Theorems 1 and 2 can be expressed explicitly [see Lemmas 1 and 2 below and corresponding results in Goldenshluger and Lepski (2011) for details].

2. We will show that for given h the expected value of the stochastic error of the estimator $\hat{f}_h$, that is, $(\mathbb{E}\|\xi_h\|_s^q)^{1/q}$, admits the upper bound of the order $O((nV_h)^{1/s-1})$ when $s \in [1,2)$ and $O((nV_h)^{-1/2})$ when $s \in (2,\infty)$. It is also obvious that

$$\mathcal{R}_s[\hat{f}_h; f] \le \|B_h\|_s + (\mathbb{E}_f\|\xi_h\|_s^q)^{1/q},$$

where $B_h(f,t) := \int K_h(t-x)f(x)\,dx - f(t)$, $t \in \mathbb{R}^d$. Thus, our estimator attains, up to a constant and remainder term, the minimum of the sum of the bias and the upper bound on the stochastic error. This form of the oracle inequality is convenient for deriving minimax and minimax adaptive results (see Section 3). Indeed, bounds on the bias and the stochastic error are usually developed separately and require completely different techniques.

3. We note that $A_{\mathcal{H}} \le O([\ln n]^d)$ and $B_{\mathcal{H}} \le O(\ln n)$ for any set $\mathcal{H} \subset [0,1]^d$ such that $h_i^{\min} \ge O(n^{-c})$, $c > 0$, $\forall i = 1,\dots,d$. If $s \in (2,\infty)$, and if the set of considered bandwidths $\mathcal{H}$ is such that $V_{\max} = [\varkappa \ln n]^{-s/2}$ for some $\varkappa > 0$, then the second term on the right-hand side of (10) and (11) can be made negligibly small by carefully choosing the constant $\varkappa$. Observe that conditions ensuring consistency of $\hat{f}_h$ are $nV_h \to \infty$ and $V_h \to 0$ as $n \to \infty$; thus the requirement $V_{\max} = [\varkappa \ln n]^{-s/2}$ is not restrictive. Note also that in the case $s \in [1,2)$ the second term on the right-hand side of (9) is exponentially small in n for any $\mathcal{H}$.

4. The condition $V_{\max} \ge 1/\sqrt{n}$ is imposed only for the sake of convenience in the presentation of our results. Clearly, we would like to have the set $\mathcal{H}$ as large as possible; hence consideration of vectors $h^{\max}$ such that $V_{\max} = V_{h^{\max}} \le 1/\sqrt{n}$ does not make much sense.

5. Note that the oracle inequalities (9), (10) and (11) of Theorems 1 and 2 hold under very mild conditions on the density f. In particular, in the case $s \in [1,2)$ the inequality (9) holds for all densities, and only boundedness of f is required for (10) and (11).

6. It should be also mentioned that if for $s \in [1,2)$ we impose additional conditions on f [e.g., such as the domination condition in Donoho et al. (1996), page 514], then the order of the stochastic error of $\hat{f}_h$ can be improved to $O((nV_h)^{-1/2})$. This will lead to the oracle inequality (9) with the term $C_1(nV_h)^{1/s-1})$ replaced by $C_1(nV_h)^{-1/2}$. However, $O((nV_h)^{1/s-1})$ is a tight upper bound on the stochastic error of $\hat{f}_h$ when no conditions on f are assumed. In particular, it is well known that smoothness condition alone is not sufficient for consistent density estimation on $\mathbb{R}^d$ with $\mathbb{L}_1$-losses [Ibragimov and Khas'minskiĭ (1981)].

2.5. $\mathbb{L}_s$*-risk oracle inequalities*. As it was mentioned above, the oracle inequalities of Theorems 1 and 2 are useful for derivation of adaptive rate optimal estimators. They are established under very mild assumptions on the density f. However, it is not clear how the second term under the infimum sign on the right-hand side of the developed oracle inequalities is compared to $\mathcal{R}_s[\hat{f}_h; f]$. Traditionally oracle inequalities compare the risk of a proposed estimator to the risk of the best estimator in the given family; cf. (2). Therefore the natural question is whether an $\mathbb{L}_s$*-risk oracle inequality* of the type (2) can be derived from the results of Theorems 1 and 2.

In this section we provide an answer to this question. We will be mostly interested in finding minimal assumptions on the underlying density f that are sufficient for establishing the $\mathbb{L}_s$-risk oracle inequality. It will be shown that this problem is directly related to establishing a lower bound on the term $(\mathbb{E}_f \|\xi_h\|_s^q)^{1/q}$.

Let $\mu \in (0,1)$ and $\nu > 0$ be fixed real numbers. Denote by $\mathbb{F}_{\mu,\nu}$ the set of all probability densities f satisfying the following condition:

$$\exists B \in \mathcal{B}(\mathbb{R}^d): \qquad \operatorname{mes}(B) \le \nu, \qquad \int_B f \ge \mu.$$

Here $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra on $\mathbb{R}^d$ and $\operatorname{mes}(\cdot)$ is the Lebesgue measure on $\mathbb{R}^d$.

Below we will assume that $f \in \mathbb{F}_{\mu,\nu}$ for some μ and ν. This condition is very weak. For example, if $\mathcal{F}$ is a set of densities such that either (i) $\mathcal{F}$ is a totally bounded subset of $\mathbb{L}_1(\mathbb{R}^d)$, or (ii) the family of probability measures $\{\mathbb{P}_f,\ f \in \mathcal{F}\}$ is tight, then for any $\mu \in (0,1)$ there exists $0 < \nu < \infty$ such that $\mathcal{F} \subseteq \mathbb{F}_{\mu,\nu}$. The statement (i) is a consequence of the Kolmogorov–Riesz compactness theorem.

THEOREM 3. *Let $s \in [2,\infty)$ and suppose that assumptions of Theorems* 1(ii) *and* 2 *are fulfilled. If $s > 2$, then assume additionally that $f \in \mathbb{F}_{\mu,\nu}$ for some μ and ν, and*

$$V_{\max} \le 2^{-1}\mu\left[\frac{\|K\|_2}{\|K\|_1}\right]^2.$$

If $n \ge C_1 = C_1(L_K, \mathrm{k}_\infty, \mathrm{f}_\infty, d, s)$, then there exists a constant $C_0 > 0$ [$C_0 = C_0(K)$ if $s = 2$ and $C_0 = C_0(K,\mu,\nu,s)$ if $s > 2$] such that

$$\mathcal{R}_s[\hat{f}; f] \le C_0 \inf_{h\in\mathcal{H}} \mathcal{R}_s[\hat{f}_h; f] + C_2 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{1/2}[\exp\{-C_3 b_{n,s}\} + \exp\{-C_4 \mathrm{f}_\infty^{-1} V_{\max}^{-2/s}\}],$$

where $b_{n,s} := n^{4/s-1}$ if $s \in (2,4)$ and $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ if $s \ge 4$. The constants C_i depend on $L_K, \mathrm{k}_\infty, d, q$ and s only.

The proof indicates that Theorem 3 follows from the fact that for any $s \in [2, \infty)$ one has

$$[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \geq c(nV_h)^{-1/2} \qquad \forall h, \tag{12}$$

where $c > 0$ is a constant. This lower bound holds under very weak conditions on the density f (for arbitrary f if $s = 2$ and $f \in \mathbb{F}_{\mu,v}$ if $s > 2$). In order to prove the similar $\mathbb{L}_s$-risk oracle inequality in the case $s \in [1, 2)$ it would be sufficient to show that $[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \geq c(nV_h)^{-1+1/s}$ for any h. However, the last lower bound cannot hold in such generality as (12). In particular, according to Remark 5 after Theorem 2, $[\mathbb{E}_f \|\xi_h\|_s^q]^{1/q} \leq c(nV_h)^{-1/2}$ for all h under a tail domination condition (e.g., for compactly supported densities). Under such a domination condition the corresponding $\mathbb{L}_s$-risk oracle inequality can be easily established using the same arguments as in the proof of Theorem 3.

2.6. *Generalization.* Although in the present paper we focus on the bandwidth selection, the proposed selection rule can be easily extended to very general families of linear estimators.

Let $\mathfrak{L}$ be the collection of functions $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\int_{\mathbb{R}^d} \mathcal{L}(t, x)\, \mathrm{d}t = 1 \qquad \forall x \in \mathbb{R}^d.$$

Consider the following family of estimators generated by $\mathfrak{L}$:

$$\mathcal{F}(\mathfrak{L}) = \left\{ \hat{f}_{\mathcal{L}}(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\cdot, X_i), \mathcal{L} \in \mathfrak{L} \right\}.$$

The objective is to propose the selection rule from the family $\mathcal{F}(\mathfrak{L})$ and to establish for the obtained estimator $\mathbb{L}_s$-oracle inequality. A close inspection of the proofs of Theorems 1 and 2 leads to the following generalization of the selection rule (8).

For any couple $\mathcal{L}, \mathcal{L}' \in \mathfrak{L}$ let

$$[\mathcal{L} \otimes \mathcal{L}'](t, x) := \int_{\mathbb{R}^d} \mathcal{L}(t, y)\mathcal{L}'(y, x)\, \mathrm{d}y$$

and define the estimator

$$\hat{f}_{\mathcal{L} \otimes \mathcal{L}'}(\cdot) = \frac{1}{n} \sum_{i=1}^n [\mathcal{L} \otimes \mathcal{L}'](\cdot, X_i).$$

Let

$$\xi_{\mathcal{L}}(t) := \frac{1}{n} \sum_{i=1}^n [\mathcal{L}(t, X_i) - \mathbb{E}_f \mathcal{L}(t, X)],$$

$$\xi_{\mathcal{L} \otimes \mathcal{L}'}(t) := \frac{1}{n} \sum_{i=1}^n \{[\mathcal{L} \otimes \mathcal{L}'](t, X_i) - \mathbb{E}_f [\mathcal{L} \otimes \mathcal{L}'](t, X)\}.$$

Suppose that for any $\mathcal{L}, \mathcal{L}' \in \mathfrak{L}$ one can find a majorant $m_s(\mathcal{L}, \mathcal{L}')$ for $\|\xi_{\mathcal{L}\otimes\mathcal{L}'} - \xi_{\mathcal{L}'}\|_s$. In other words, suppose that the expectation

$$\mathbb{E}_f \sup_{(\mathcal{L},\mathcal{L}')\in\mathfrak{L}\times\mathfrak{L}} [\|\xi_{\mathcal{L}\otimes\mathcal{L}'} - \xi_{\mathcal{L}'}\|_s - m_s(\mathcal{L}, \mathcal{L}')]_+^q$$

is "small," and analogues of Lemmas 1 and 2 given below are proved. We refer to Goldenshluger and Lepski (2011), where results of this type for various collections $\mathfrak{L}$ can be found.

For every $\mathcal{L} \in \mathfrak{L}$ let

$$\hat{R}_{\mathcal{L}} := \sup_{\mathcal{L}'\in\mathfrak{L}} [\|\hat{f}_{\mathcal{L}\otimes\mathcal{L}'} - \hat{f}_{\mathcal{L}'}\|_s - m_s(\mathcal{L}, \mathcal{L}')]_+ + \sup_{\mathcal{L}'\in\mathfrak{L}} m_s(\mathcal{L}', \mathcal{L}), \tag{13}$$

and define

$$\widehat{\mathcal{L}} := \arg\inf_{\mathcal{L}\in\mathfrak{L}} \hat{R}_{\mathcal{L}}. \tag{14}$$

The selected estimator is $\hat{f} = \hat{f}_{\widehat{\mathcal{L}}}$.

In order to prove analogues of Theorems 1 and 2 the following assumption (*commutativity property*) on the collection $\mathfrak{L}$ has to be imposed:

$$\int_{\mathbb{R}^d} \mathcal{L}(\cdot, y)\mathcal{L}'(y, \cdot)\,\mathrm{d}y = \int_{\mathbb{R}^d} \mathcal{L}'(\cdot, y)\mathcal{L}(y, \cdot)\,\mathrm{d}y \qquad \forall \mathcal{L}, \mathcal{L}' \in \mathfrak{L}. \tag{15}$$

Thus, using the commutativity property (15) and majorants for the $\mathbb{L}_s$-norms of empirical processes derived in Goldenshluger and Lepski (2011), one can establish $\mathbb{L}_s$-oracle inequalities for the selection rule (13)–(14).

3. Adaptive estimation of densities with anisotropic smoothness. In this section we illustrate the use of oracle inequalities of Theorems 1 and 2 for derivation of adaptive rate optimal density estimators.

We start with the definition of the *anisotropic Nikol'skii class of functions*.

DEFINITION 1. Let $p \in [1, \infty]$, $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_i > 0$, and $L > 0$. We say that a density $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the anisotropic Nikol'skii class $N_{p,d}(\alpha, L)$ of functions if:

(i) $\|D_i^{\lfloor\alpha_i\rfloor} f\|_p \le L$, for all $i = 1, \ldots, d$;

(ii) for all $i = 1, \ldots, d$, and all $z \in \mathbb{R}^1$

$$\left\{\int |D_i^{\lfloor\alpha_i\rfloor} f(t_1, \ldots, t_i + z, \ldots, t_d) - D_i^{\lfloor\alpha_i\rfloor} f(t_1, \ldots, t_i, \ldots, t_d)|^p\,\mathrm{d}t\right\}^{1/p} \le L|z|^{\alpha_i - \lfloor\alpha_i\rfloor}.$$

Here $D_i^k f$ denotes the kth-order partial derivative of f with respect to the variable t_i and $\lfloor\alpha_i\rfloor$ is the largest integer strictly less than α_i.

The functional classes $N_{p,d}(\alpha, L)$ were considered in approximation theory by Nikol'skii; see, for example, Nikol'skiĭ (1969). Minimax estimation of densities from the class $N_{p,d}(\alpha, L)$ was considered in Ibragimov and Khas'minskiĭ (1981). We refer also to Kerkyacharian, Lepski and Picard (2001) where the problem of adaptive estimation over a scale of classes $N_{p,d}(\alpha, L)$ was treated for the Gaussian white noise model.

Consider the following family of kernel estimators. Let u be an integrable, compactly supported function on $\mathbb{R}$ such that $\int u(y)\,dy = 1$. As in Kerkyacharian, Lepski and Picard (2001), for some integer number l we put

$$u_l(y) := \sum_{k=1}^{l} \binom{l}{k} (-1)^{k+1} \frac{1}{k} u\left(\frac{y}{k}\right),$$

and define

$$K(t) := \prod_{i=1}^{d} u_l(t_i), \qquad t = (t_1, \dots, t_d). \tag{16}$$

The kernel K constructed in this way is bounded and compactly supported, and it is easily verified that

$$\int K(t)\,dt = 1, \qquad \int K(t) t^k\,dt = 0 \qquad \forall |k| = 1, \dots, l-1,$$

where $k = (k_1, \dots, k_d)$ is the multi-index, $k_i \geq 0$, $|k| = k_1 + \cdots + k_d$ and $t^k = t_1^{k_1} \cdots t_d^{k_d}$ for $t = (t_1, \dots, t_d)$.

For fixed $\alpha = (\alpha_1, \dots, \alpha_d)$ set $1/\bar{\alpha} = \sum_{i=1}^{d} (1/\alpha_i)$ and define

$$\varphi_{n,s}(\bar{\alpha}) := L^{-\gamma_s/(\bar{\alpha}+\gamma_s)} n^{-\gamma_s \bar{\alpha}/(\bar{\alpha}+\gamma_s)}, \qquad \gamma_s := \begin{cases} 1 - 1/s, & s \in (1, 2], \\ 1/2, & s \in (2, \infty). \end{cases}$$

THEOREM 4. *Let $\mathcal{F}(\mathcal{H})$ be the family of kernel estimators defined in (1), (3) and (4) that is associated with the kernel (16). Let $\hat{f}$ denote the estimator given by selection according to our rule (6)–(8) from the family $\mathcal{F}(\mathcal{H})$.*

(i) *Let $s \in (1, 2)$, and assume that $h_i^{\min} = 1/n$ and $h_i^{\max} = 1$, $\forall i = 1, \dots, d$. Then for any class $N_{s,d}(\alpha, L)$ such that $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$ one has*

$$\limsup_{n \to \infty} \{[\varphi_{n,s}(\bar{\alpha})]^{-1} \mathcal{R}_s[\hat{f}; N_{s,d}(\alpha, L)]\} < \infty.$$

(ii) *Let $s \in [2, \infty)$, and assume that $h_i^{\min} = \varkappa_1/n$ and $h_i^{\max} = [\varkappa_2 \ln n]^{-s/(2d)}$, $\forall i = 1, \dots, d$ for some constants $\varkappa_1$ and $\varkappa_2$. Then for any class $N_{s,d}(\alpha, L)$ such that $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$ one has*

$$\limsup_{n \to \infty} \{[\varphi_{n,s}(\bar{\alpha})]^{-1} \mathcal{R}_s[\hat{f}; N_{s,d}(\alpha, L)]\} < \infty.$$

It is well known that $\varphi_{n,s}(\bar{\alpha})$ is the minimax rate of convergence in estimation of densities from the class $N_{s,d}(\alpha, L)$ [see Ibragimov and Khas'minskiĭ (1981) and Hasminskii and Ibragimov (1990)]. Therefore Theorem 4 shows that our estimator $\hat{f}$ is adaptive minimax over a scale of the classes $N_{s,d}(\alpha, L)$ indexed by α and L.

The above result holds when both the smoothness and the accuracy are measured in the same $\mathbb{L}_s$-norm. We demonstrate below that if the additional condition of compact support is imposed, then the resulting estimator is adaptive minimax over a much larger scale of functional classes.

DEFINITION 2. Let $p \in [1, \infty]$, $\alpha = (\alpha_1, \dots, \alpha_d)$, $\alpha_i > 0$, $L > 0$, and let Q be a fixed cube in $\mathbb{R}^d$. We say that a density $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the functional class $W_{p,d}(\alpha, L, Q)$ if $f \in N_{p,d}(\alpha, L)$, and $\operatorname{supp}(f) \subseteq Q$.

THEOREM 5. *Let $s \in [1, \infty)$, and assume that $h_i^{\min} = \varkappa_1/n$ and $h_i^{\max} = [\varkappa_2 \ln n]^{-[s\vee 2]/(2d)}$, $\forall i = 1, \dots, d$ for some constants $\varkappa_1$ and $\varkappa_2$. Let $\mathcal{F}(\mathcal{H})$ be the corresponding family of kernel estimators that is associated with the kernel (16). Let $\hat{f}$ denote the estimator given by the selection procedure (6)–(8) with s substituted by $s \vee 2$. Then for any class $W_{p,d}(\alpha, L, Q)$ such that $p \geq [s \vee 2]$, $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$*

$$\limsup_{n\to\infty} \{[\psi_{n,s}(\bar{\alpha})]^{-1} \mathcal{R}_s[\hat{f}; W_{p,d}(\alpha, L, Q)]\} < \infty,$$

where

$$\psi_{n,s}(\bar{\alpha}) := \left(L[\operatorname{mes}\{Q\}]^{(p-[s\vee 2])/p[s\vee 2]}\right)^{1/(2\bar{\alpha}+1)} n^{-\bar{\alpha}/(2\bar{\alpha}+1)}.$$

Theorem 5 shows that if $s \in [1, \infty)$, then the estimator $\hat{f}$ given by our selection procedure achieves the minimax rate of convergence simultaneously on every class $W_{p,d}(\alpha, L, Q)$ with any $p \geq [s \vee 2]$, $\max_{i=1,\dots,d} \lfloor \alpha_i \rfloor \leq l - 1$, $L > 0$ and any fixed support Q. It should be especially stressed that no information about the support set Q and the index p are used in construction of $\hat{f}$.

4. Proofs. First we recall that the accuracy of estimators $\hat{f}_h$ and $\hat{f}_{h,\eta}$, $h, \eta \in \mathcal{H}$, is characterized by the bias and stochastic error given by

$$B_h(f, t) := \int K_h(t - x) f(x)\, dx - f(t),$$

$$\xi_h(t) := \frac{1}{n} \sum_{i=1}^{n} [K_h(t - X_i) - \mathbb{E}_f K_h(t - X)]$$

and

$$B_{h,\eta}(f, t) := \int [K_h * K_\eta](t - x) f(x)\, dx - f(t),$$

$$\xi_{h,\eta}(t) := \frac{1}{n} \sum_{i=1}^{n} \{[K_h * K_\eta](t - X_i) - \mathbb{E}_f [K_h * K_\eta](t - X)\},$$

respectively.

The proofs extensively use results from Goldenshluger and Lepski (2011); in what follows for the sake of brevity we refer to this paper as GL (2011).

4.1. *Auxiliary results.* We start with two auxiliary lemmas that establish probability and moment bounds on $\mathbb{L}_s$-norms of the processes ξ_h and $\xi_{h,\eta}$. Proofs of these results are given in the Appendix.

LEMMA 1. *Let Assumptions* (K1) *and* (K2) *hold.*

(i) *If* $s \in [1,2)$, *then for all* $n \geq 4^{2s/(2-s)}$ *one has*

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}}[\|\xi_h\|_s - 32\rho_s(K_h)]_+^q\Big\}^{1/q}\\ &\qquad\leq \delta_{n,s}^{(1)} := C_1 A_{\mathcal{H}}^{2/q} n^{1/s}\exp\Big\{-\frac{2n^{2/s-1}}{37q}\Big\},\end{aligned}\tag{17}$$

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}}[\|\xi_{h,\eta}\|_s - 32\rho_s(K_h * K_\eta)]_+^q\Big\}^{1/q}\\ &\qquad\leq \delta_{n,s}^{(2)} := C_2 A_{\mathcal{H}}^{4/q} n^{1/s}\exp\Big\{-\frac{2n^{2/s-1}}{37q}\Big\}.\end{aligned}\tag{18}$$

(ii) *Let* $f \in \mathbb{F}$, *and assume that* $8[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}] \leq 1$; *then for all* $f \in \mathbb{F}$ *one has*

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}}\Big[\|\xi_h\|_2 - \frac{25}{3}\rho_2(K_h)\Big]_+^q\Big\}^{1/q}\\ &\qquad\leq \delta_{n,2}^{(1)} := C_3 A_{\mathcal{H}}^{2/q} n^{1/2}\exp\Big\{-\frac{1}{16q[V_{\max}\mathrm{f}_\infty^2 + 4n^{-1/2}]}\Big\},\end{aligned}\tag{19}$$

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}}\Big[\|\xi_{h,\eta}\|_2 - \frac{25}{3}\rho_2(K_h * K_\eta)\Big]_+^q\Big\}^{1/q}\\ &\qquad\leq \delta_{n,2}^{(2)} := C_4 A_{\mathcal{H}}^{4/q} n^{1/2}\exp\Big\{-\frac{1}{16q[\mathrm{f}_\infty^2 V_{\max} + 4n^{-1/2}]}\Big\}.\end{aligned}\tag{20}$$

The constants C_i, $i = 1,\dots,4$, *depend on* L_K, k_∞, d *and* q *only.*

LEMMA 2. *Let Assumptions* (K1) *and* (K2) *hold,* $f \in \mathbb{F}$, $s > 2$, *and assume that*

$$n \geq C_1, \qquad nV_{\min} > C_2, \qquad V_{\max} \geq 1/\sqrt{n}.$$

Then the following statements hold:

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}}[\|\xi_h\|_s - 32\hat{r}_s(K_h)]_+^q\Big\}^{1/q}\\ &\qquad\le \delta_{n,s}^{(1)} := C_3 A_{\mathcal{H}}^{2/q} B_{\mathcal{H}}^{1/q} n^{1/2}\exp\Big\{-\frac{C_4}{\mathrm{f}_\infty V_{\max}^{2/s}}\Big\},\end{aligned}\tag{21}$$

$$\begin{aligned}&\Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}}[\|\xi_{h,\eta}\|_s - 32\hat{r}_s(K_h * K_\eta)]_+^q\Big\}^{1/q}\\ &\qquad\le \delta_{n,s}^{(2)} := C_5 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{1/2}\exp\Big\{-\frac{C_6}{\mathrm{f}_\infty V_{\max}^{2/s}}\Big\}.\end{aligned}\tag{22}$$

In addition, for any $H_1 \subseteq \mathcal{H}$ *and* $H_2 \subseteq \mathcal{H}$

$$\begin{aligned}\mathbb{E}_f \sup_{h\in H_1}[\hat{r}_s(K_h)]^q &\le (1+8D_s)^q \sup_{h\in H_1}[r_s(K_h)]^q\\ &\quad + C_7 A_{\mathcal{H}}^2 B_{\mathcal{H}} n^{q(s-2)/(2s)}\exp\{-C_8 b_{n,s}\},\end{aligned}\tag{23}$$

$$\begin{aligned}\mathbb{E}_f \sup_{(h,\eta)\in H_1\times H_2}[\hat{r}_s(K_h * K_\eta)]^q &\le (1+8D_s)^q \sup_{(h,\eta)\in H_1\times H_2}[r_s(K_h * K_\eta)]^q\\ &\quad + C_9 A_{\mathcal{H}}^4 B_{\mathcal{H}} n^{q(s-2)/(2s)}\exp\{-C_{10} b_{n,s}\},\end{aligned}\tag{24}$$

where $b_{n,s} := n^{4/s-1}$ *if* $s \in (2,4)$ *and* $b_{n,s} := [\mathrm{f}_\infty V_{\max}^{4/s}]^{-1}$ *if* $s \in [4,\infty)$. *The constants* C_i, $i = 2,\dots,10$, *depend on* L_K, k_∞, d, q *and* s *only, while* C_1 *depends also on* f_∞.

4.2. *Proofs of Theorems* 1 *and* 2. The proofs of both theorems (which we break into several steps) follow along the same lines.

We note that in the case $s \in [2,\infty)$ the condition $f \in \mathbb{F}$ implies that $f \in \mathbb{L}_s(\mathbb{R}^d)$. If $s \in (1,2)$, then by Assumptions (K1) and (K2), we have that $\mathbb{P}_f\{\hat{f}_h \in \mathbb{L}_s(\mathbb{R}^d)\} = 1$ for any $\mathcal{X}_n$-measurable vector $h \in \mathcal{H}$ and for any n. Hence, if $f \notin \mathbb{L}_s(\mathbb{R}^d)$, then $\mathcal{R}[\hat{f}_h; f] = +\infty$, $\forall h \in \mathcal{H}$, and the result (i) of Theorem 1 holds trivially. Thus, we can assume that $f \in \mathbb{L}_s(\mathbb{R}^d)$ when $s \in (1,2)$.

1°. First we show that for any $h, \eta \in \mathcal{H}$

$$B_{h,\eta}(f,x) = B_\eta(f,x) + \int K_\eta(y-x)B_h(f,y)\,\mathrm{d}y\tag{25}$$

$$= B_h(f,x) + \int K_h(y-x)B_\eta(f,y)\,\mathrm{d}y.\tag{26}$$

Indeed, by the Fubini theorem,

$$\begin{aligned}&\int [K_h * K_\eta](t-x)f(t)\,\mathrm{d}t\\ &\qquad = \int\Big[\int K_h(t-y)K_\eta(y-x)\,\mathrm{d}y\Big]f(t)\,\mathrm{d}t\end{aligned}$$

$$= \int \left[\int K_h(t-y) f(t)\,\mathrm{d}t - f(y) \right] K_\eta(y-x)\,\mathrm{d}y + \int K_\eta(y-x) f(y)\,\mathrm{d}y$$
$$= \int K_\eta(y-x) f(y)\,\mathrm{d}y + \int K_\eta(y-x) B_h(f,y)\,\mathrm{d}y.$$

Subtracting $f(x)$ from both sides of the last equality we come to (25); (26) follows similarly.

2°. Let $m_s(\cdot,\cdot)$ and $m_s^*(\cdot)$ be given by (6), and define

$$\delta_{n,s} := \Big\{ \mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} [\|\hat\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)]_+^q \Big\}^{1/q}. \tag{27}$$

Let $\hat f = \hat f_{\hat h}$ be the estimator defined in (7)–(8). Our first goal is to prove that

$$\mathcal{R}_s[\hat f; f] \le \inf_{h\in\mathcal{H}} \{(1+3\|K\|_1)\mathcal{R}_s[\hat f_h; f] + 3(\mathbb{E}_f[m_s^*(h)]^q)^{1/q}\} + 3\delta_{n,s}. \tag{28}$$

By the triangle inequality for any $\eta \in \mathcal{H}$

$$\|\hat f_{\hat h} - f\|_s \le \|\hat f_{\hat h} - \hat f_{\hat h,\eta}\|_s + \|\hat f_{\hat h,\eta} - \hat f_\eta\|_s + \|\hat f_\eta - f\|_s, \tag{29}$$

and we are going to bound the first two terms on the right-hand side.

Define

$$\bar B_h(f) := \sup_{\eta\in\mathcal{H}} \left\| \int K_\eta(t-\cdot) B_h(f,t)\,\mathrm{d}t \right\|_s, \qquad h \in \mathcal{H}.$$

We have for any $h \in \mathcal{H}$

$$\begin{aligned}
\hat R_h - m_s^*(h) &= \sup_{\eta\in\mathcal{H}} [\|\hat f_{h,\eta} - \hat f_\eta\|_s - m_s(h,\eta)] \\
&\le \sup_{\eta\in\mathcal{H}} [\|B_{h,\eta}(f,\cdot) - B_\eta(f,\cdot)\|_s + \|\hat\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)] \\
&\le \bar B_h(f) + \sup_{\eta\in\mathcal{H}} [\|\hat\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)]_+ =: \bar B_h(f) + \zeta.
\end{aligned}$$

Here the second line is by the triangle inequality and the third line is by (25) and definition of $\bar B_h(f)$. Therefore for any $h \in \mathcal{H}$ one has

$$\hat R_h \le \bar B_h(f) + m_s^*(h) + \zeta. \tag{30}$$

By (26) for any $h, \eta \in \mathcal{H}$

$$\begin{aligned}
\|\hat f_{h,\eta} - \hat f_h\|_s &\le \|B_{h,\eta}(f,\cdot) - B_h(f,\cdot)\|_s + \|\hat\xi_{h,\eta} - \xi_h\|_s \\
&\le \bar B_\eta(f) + \zeta + \sup_{\eta\in\mathcal{H}} m_s(\eta,h) \\
&= \bar B_\eta(f) + m_s^*(h) + \zeta \le \bar B_\eta(f) + \hat R_h + \zeta,
\end{aligned}$$

where the last inequality is by definition of $\hat{R}_h$. In particular, letting $h = \hat{h}$ we have that for any $\eta \in \mathcal{H}$

$$\begin{aligned} \|\hat{f}_{\hat{h},\eta} - \hat{f}_{\hat{h}}\|_s &\le \bar{B}_\eta(f) + \hat{R}_{\hat{h}} + \zeta \\ &\le \bar{B}_\eta(f) + \hat{R}_\eta + \zeta \le 2\bar{B}_\eta(f) + m_s^*(\eta) + 2\zeta, \end{aligned} \tag{31}$$

where we have used that $\hat{R}_{\hat{h}} \le \hat{R}_\eta$, $\forall \eta \in \mathcal{H}$ and (30).

Furthermore, for any $\eta \in \mathcal{H}$

$$\begin{aligned} \|\hat{f}_{\hat{h},\eta} - \hat{f}_\eta\|_s &= \|\hat{f}_{\hat{h},\eta} - \hat{f}_\eta\|_s - m_s(\hat{h}, \eta) + m_s(\hat{h}, \eta) \\ &\le \hat{R}_{\hat{h}} + m_s^*(\eta) \le \hat{R}_\eta + m_s^*(\eta) \le \bar{B}_\eta(f) + 2m_s^*(\eta) + \zeta, \end{aligned} \tag{32}$$

where the first inequality is by definition of $\hat{R}_h$ and $m_s^*(\cdot)$, the second inequality holds by definition of $\hat{h}$, and the last inequality follows from (30).

Combining (29), (31) and (32) we get for any $\eta \in \mathcal{H}$ that

$$\begin{aligned} \|\hat{f}_{\hat{h}} - f\|_s &\le \|\hat{f}_{\hat{h}} - \hat{f}_{\hat{h},\eta}\|_s + \|\hat{f}_{\hat{h},\eta} - \hat{f}_\eta\|_s + \|\hat{f}_\eta - f\|_s \\ &\le \|\hat{f}_\eta - f\|_s + 3\bar{B}_\eta(f) + 3m_s^*(\eta) + 3\zeta. \end{aligned}$$

Taking this expression to the power q, computing the expectation and using the fact that $[\mathbb{E}_f|\zeta|^q]^{1/q} = \delta_{n,s}$ we obtain

$$\mathcal{R}_s[\hat{f}; f] \le \inf_{h\in\mathcal{H}} \{\mathcal{R}_s[\hat{f}_h; f] + 3\bar{B}_h(f) + 3(\mathbb{E}_f[m_s^*(h)]^q)^{1/q}\} + 3\delta_{n,s}. \tag{33}$$

By the Young inequality

$$\bar{B}_h(f) \le \Big(\sup_{\eta\in\mathcal{H}} \|K_\eta\|_1\Big)\|B_h(f,\cdot)\|_s = \|K\|_1\|B_h(f,\cdot)\|_s.$$

In addition [see (39)–(40)],

$$\|B_h(f,\cdot)\|_s \le \mathcal{R}_s[\hat{f}_h; f] \qquad \forall h \in \mathcal{H}.$$

Combining this with (33), we complete the proof of (28).

3°. Lemmas 1 and 2 lead to an upper bound on the quantity $\delta_{n,s}$ given in (27). Indeed, by definition of $m_s(\cdot,\cdot)$ [see (6)] we have

$$\begin{aligned} \delta_{n,s} &= \Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} [\|\xi_{h,\eta} - \xi_\eta\|_s - m_s(h,\eta)]_+^q\Big\}^{1/q} \\ &\le \Big\{\mathbb{E}_f \sup_{(h,\eta)\in\mathcal{H}\times\mathcal{H}} [\|\xi_{h,\eta}\|_s - g_s(K_h * K_\eta)]_+^q\Big\}^{1/q} \\ &\quad + \Big\{\mathbb{E}_f \sup_{h\in\mathcal{H}} [\|\xi_h\|_s - g_s(K_h)]_+^q\Big\}^{1/q} \le \delta_{n,s}^{(1)} + \delta_{n,s}^{(2)}, \end{aligned} \tag{34}$$

where expressions for $\delta_{n,s}^{(1)}$ and $\delta_{n,s}^{(2)}$ depending on the value of $s \in [1,\infty)$ are given in (17)–(18), (19)–(20) and (21)–(22).

In order to apply (28) it remains to bound $\{\mathbb{E}_f[m_s^*(h)]^q\}^{1/q}$.

4°. We start with the case $s \in [1,2)$. Here, by definition,

$$m_s^*(h) = \sup_{\eta\in\mathcal{H}} m_s(\eta,h) = g_s(K_h) + \sup_{\eta\in\mathcal{H}} g_s(K_\eta * K_h)$$
$$= 128n^{1/s-1}\Big(\|K_h\|_s + \sup_{\eta\in\mathcal{H}}\|K_h * K_\eta\|_s\Big) \le 128[1+\|K\|_1]\mathrm{k}_\infty(nV_h)^{1/s-1}.$$

Therefore applying (28), and taking into account (34), (17) and (18), we come to the statement (i) of Theorem 1.

The statement (ii) of Theorem 1 dealing with the case $s=2$ follows similarly by application of (28) and (34), (19) and (20). This completes the proof of Theorem 1.

5°. Now consider the case $s \in (2,\infty)$. Because

$$\begin{aligned} m_s^*(h) &= \sup_{\eta\in\mathcal{H}} m_s(\eta,h) = g_s(K_h) + \sup_{\eta\in\mathcal{H}} g_s(K_\eta * K_h) \\ &= 32\hat{r}_s(K_h) + 32\sup_{\eta\in\mathcal{H}}\hat{r}_s(K_\eta * K_h), \end{aligned} \tag{35}$$

it suffices to bound from above $[\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q}$ and $[\mathbb{E}_f \sup_{\eta\in\mathcal{H}}|\hat{r}_s(K_h * K_\eta)|^q]^{1/q}$. Using (23) of Lemma 2 with $H_1=\{h\}$ we have

$$[\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q} \le c_1 r_s(K_h) + c_2 A_{\mathcal{H}}^{2/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)} \exp\{-c_3 b_{n,s}\}.$$

In addition, by the Young inequality,

$$\begin{aligned} \rho_s(K_h) &= D_s n^{-1/2}\|K_h^2 * f\|_{s/2}^{1/2} + n^{1/s-1}\|K_h\|_s \\ &\le D_s n^{-1/2}\|K_h\|_2\big\|\sqrt{f}\big\|_s + (nV_h)^{-1+1/s}\|K\|_s \\ &\le D_s \mathrm{f}_\infty^{1/2}\|K\|_2(nV_h)^{-1/2} + \|K\|_s(nV_h)^{-1+1/s} \le c_4 \mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}; \end{aligned}$$

here we have used that $\|\sqrt{f}\|_s = (\int f^{s/2}(x)\,dx)^{1/s} \le (\mathrm{f}_\infty^{s/2-1}\int f(x)\,dx)^{1/s} \le \mathrm{f}_\infty^{1/2}$. Hence

$$\begin{aligned} [\mathbb{E}_f|\hat{r}_s(K_h)|^q]^{1/q} &\le c_5 \mathrm{f}_\infty^{1/2}(nV_h)^{-1/2} \\ &\quad + c_2 A_{\mathcal{H}}^{2/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)} \exp\{-c_3 b_{n,s}\}. \end{aligned} \tag{36}$$

Now, applying (24) with $H_1=\{h\}$ and $H_2=\mathcal{H}$ we obtain

$$\begin{aligned} \Big[\mathbb{E}_f \sup_{\eta\in\mathcal{H}}|\hat{r}_s(K_h * K_\eta)|^q\Big]^{1/q} &\le c_6 \sup_{\eta\in\mathcal{H}} r_s(K_h * K_\eta) \\ &\quad + c_7 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)} \exp\{-c_8 b_{n,s}\}. \end{aligned}$$

In addition, similar to the above,

$$\begin{aligned} \sup_{\eta\in\mathcal{H}} \rho_s(K_h * K_\eta) &\le \sup_{\eta\in\mathcal{H}}\big\{D_s n^{-1/2}\|K_h * K_\eta\|_2\big\|\sqrt{f}\big\|_s + n^{-1+1/s}\|K_h * K_\eta\|_s\big\} \\ &\le c_8 \mathrm{f}_\infty^{1/2} \sup_{\eta\in\mathcal{H}}[n(V_h \vee V_\eta)]^{1/2} \le c_9 \mathrm{f}_\infty^{1/2}(nV_h)^{-1/2}. \end{aligned}$$

Therefore the last two bounds yield

$$\Big[\mathbb{E}_f \sup_{\eta\in\mathcal{H}} |\hat{r}_s(K_h * K_\eta)|^q\Big]^{1/q} \le c_{10}\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2} + c_7 A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)} \exp\{-c_8 b_{n,s}\}.$$

This along with (36) and (35) results in

$$[\mathbb{E}_f |m_s^*(K_h)|^q]^{1/q} \le c_{11}\mathrm{f}_\infty^{1/2}(nV_h)^{-1/2} + c_{12} A_{\mathcal{H}}^{4/q} B_{\mathcal{H}}^{1/q} n^{(s-2)/(2s)} \exp\{-c_{13} b_{n,s}\}.$$

Combining this bound with (21), (22) and (34), and applying (28), we complete the proof of Theorem 2.

4.3. *Proof of Theorem 3.* Throughout the proof we denote by $c_0, c_1, \dots,$ the positive constants depending only on the kernel K, the index s and the quantity f_∞. We divide the proof into four steps.

1°. Let us prove that for any $q \ge 1$ and $h \in \mathcal{H}$

$$3\mathcal{R}_s[\hat{f}_h; f] \ge \|B_h(f,\cdot)\|_s + \mathbb{E}_f \|\xi_h\|_s. \tag{37}$$

Indeed, in view of the Jensen inequality for any $q \ge 1$

$$\mathcal{R}_s[\hat{f}_h; f] \ge \mathbb{E}_f \|\hat{f}_h - f\|_s = \mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s. \tag{38}$$

Denote by $\mathbb{B}_p(1)$, $1 \le p \le \infty$, the unit ball in $\mathbb{L}_p(\mathbb{R}^d)$. By the duality argument

$$\mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s = \mathbb{E}_f \sup_{\ell\in\mathbb{B}_r(1)} \int \ell(t)[B_h(f,t) + \xi_h(t)]\,\mathrm{d}t, \qquad r = \frac{s}{s-1}.$$

Let $\ell_0 \in \mathbb{B}_r(1)$ be such that $\|B_h(f,\cdot)\|_s = \int \ell_0(t) B_h(f,t)\,\mathrm{d}t$; then

$$\mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s \ge \mathbb{E}_f \int \ell_0(t)[B_h(f,t) + \xi_h(t)]\,\mathrm{d}t = \|B_h(f,\cdot)\|_s. \tag{39}$$

Here we have used that $\mathbb{E}_f \xi_h(t) = 0$, $\forall t \in \mathbb{R}^d$. We also have by the triangle inequality

$$\mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s \ge \mathbb{E}_f \|\xi_h\|_s - \|B_h(f,\cdot)\|_s. \tag{40}$$

Summing up the inequalities in (39) and (40) we get

$$\mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s \ge 2^{-1}\mathbb{E}_f \|\xi_h\|_s. \tag{41}$$

Thus, in view of (39) and (41) for any $\alpha \in (0,1)$

$$\mathbb{E}_f \|B_h(f,\cdot) + \xi_h\|_s \ge (1-\alpha)\|B_h(f,\cdot)\|_s + 2^{-1}\alpha \mathbb{E}_f \|\xi_h\|_s. \tag{42}$$

Choosing $\alpha = 2/3$, we arrive to (37) in view of (38).

In view of (37), the assertion of the theorem will follow from the statement of Theorem 2 if we show that

$$\mathbb{E}_f\|\xi_h\|_s \ge c_0(nV_h)^{-1/2}.$$

2°. Let $b>0$ be a constant to be specified, and put $a=b^{-1}\sqrt{nV_h}$. By duality

$$\mathbb{E}_f\|\xi_h\|_s = \mathbb{E}_f \sup_{\ell\in\mathbb{B}_r(1)} \int \ell(t)\xi_h(t)\,\mathrm{d}t, \qquad r=\frac{s}{s-1}. \tag{43}$$

Define the random event $\mathcal{A}=\{a\xi_h\in\mathbb{B}_2(1)\}$, and note that if $\mathcal{A}$ occurs, then by the Hölder inequality

$$ag\xi_h\in\mathbb{B}_r(1) \qquad \forall g\in\mathbb{B}_{2r/(2-r)}(1). \tag{44}$$

Recall that $s\ge 2$ implies $r\in[1,2]$, and if $r=s=2$, then we formally put $\frac{2r}{2-r}=\infty$.

If the event $\mathcal{A}$ occurs, then $\mathbb{B}_r(1)\supseteq\{ag\xi_h : g\in\mathbb{B}_{2r/(2-r)}(1)\}$. Therefore, by (43) and (44)

$$\begin{aligned}
\mathbb{E}_f\|\xi_h\|_s &\ge a\mathbb{E}_f\biggl[\mathbb{I}(\mathcal{A})\sup_{g\in\mathbb{B}_{2r/(2-r)}(1)}\int g(t)\xi_h^2(t)\,\mathrm{d}t\biggr]\\
&\ge a\sup_{g\in\mathbb{B}_{2r/(2-r)}(1)}\mathbb{E}_f\biggl[\mathbb{I}(\mathcal{A})\int g(t)\xi_h^2(t)\,\mathrm{d}t\biggr]\\
&= a\sup_{g\in\mathbb{B}_{2r/(2-r)}(1)}\int g(t)[\mathbb{E}_f\mathbb{I}(\mathcal{A})\xi_h^2(t)]\,\mathrm{d}t = a\|\mathbb{E}_f\xi_h^2(\cdot)\mathbb{I}(\mathcal{A})\|_{2s/(s+2)}\\
&\ge a\bigl[\|\mathbb{E}_f\xi_h^2(\cdot)\|_{2s/(s+2)} - \|\mathbb{E}_f\xi_h^2(\cdot)\mathbb{I}(\bar{\mathcal{A}})\|_{2s/(s+2)}\bigr],
\end{aligned} \tag{45}$$

where $\bar{\mathcal{A}}$ is the event complementary to $\mathcal{A}$.

Now consider separately two cases: $s=2$ and $s>2$.

3°. If $s=2$, we get from (45)

$$\mathbb{E}_f\|\xi_h\|_2 \ge a\biggl[\int\mathbb{E}_f\xi_h^2(t)\,\mathrm{d}t - \mathbb{E}_f\bigl\{\|\xi_h\|_2^2\mathbb{I}\bigl(\|\xi_h\|_2\ge b(nV_h)^{-1/2}\bigr)\bigr\}\biggr]. \tag{46}$$

Note that

$$\mathbb{E}_f\xi_h^2(t) = n^{-1}\int K_h^2(t-x)f(x)\,\mathrm{d}x - n^{-1}\biggl[\int K_h(t-x)f(x)\,\mathrm{d}x\biggr]^2 \tag{47}$$

and, therefore,

$$\int\mathbb{E}_f\xi_h^2(t)\,\mathrm{d}t = \frac{\|K\|_2^2}{nV_h} - n^{-1}\int\biggl[\int K_h(t-x)f(x)\,\mathrm{d}x\biggr]^2\mathrm{d}t.$$

The Young inequality yields

$$\int\biggl[\int K_h(t-x)f(x)\,\mathrm{d}x\biggr]^2\mathrm{d}t \le \|K_h\|_1^2\|f\|_2^2 \le \|K\|_1^2\mathrm{f}_\infty. \tag{48}$$

Here we have used that $f \in \mathbb{F}$. Thus, in view of $V_h \le V_{\max} \le 1/8$ [see assumption of part (ii) of Theorem 1], we obtain

$$\int \mathbb{E}_f \xi_h^2(t)\,\mathrm{d}t \ge \frac{\|K\|_2^2}{nV_h} - \frac{\|K\|_1^2 \mathrm{f}_\infty}{n} \ge c_1 (nV_h)^{-1}. \tag{49}$$

It follows from Theorem 1 of GL (2011) that for any $x \ge 2$

$$\mathbb{P}\left\{ \|\xi_h\|_2 \ge \frac{x\|K\|_2}{\sqrt{nV_h}} \right\} \le e^{c_2(1-x)} \tag{50}$$

and, therefore, putting $b = y\|K\|_2$, $y \ge 2$, we obtain

$$\mathbb{E}_f \left\{ \|\xi_h\|_2^2 \mathbb{I}\left(\|\xi_h\|_2 \ge \frac{y\|K\|_2}{\sqrt{nV_h}} \right) \right\} \le 2\|K\|_2^2 (nV_h)^{-1} \int_y^\infty x e^{c_2(1-x)}\,\mathrm{d}x. \tag{51}$$

Choosing y sufficiently large in order to make the latter integral less than $\frac{c_1}{4\|K\|_2^2}$, we obtain from (46), (49) and (51)

$$\mathbb{E}_f \|\xi_h\|_2 \ge c_3 (nV_h)^{-1/2}.$$

The theorem is proved in the case $s = 2$.

4°. Return now to the case $s > 2$. Note first that

$$\begin{aligned} \|\mathbb{E}_f \xi_h^2(\cdot)\|_{2s/(s+2)} &\ge \left(\int_B |\mathbb{E}_f \xi_h^2(t)|^{2s/(s+2)}\,\mathrm{d}t \right)^{(s+2)/(2s)} \\ &\ge \nu^{(2-s)/(2s)} \int_B \mathbb{E}_f \xi_h^2(t)\,\mathrm{d}t. \end{aligned} \tag{52}$$

The last relation is obtained by the reversed Hölder inequality. Taking into account that $\int_B f(t)\,\mathrm{d}t \ge \mu$, we get, using (47) and (48),

$$\int_B \mathbb{E}_f \xi_h^2(t)\,\mathrm{d}t \ge \frac{\mu\|K\|_2^2}{nV_h} - \frac{\|K\|_1^2 \mathrm{f}_\infty}{n} \ge c_4 \mu (nV_h)^{-1}. \tag{53}$$

Here we have used that $V_h \le 2^{-1}\mu\|K\|_2^2/\|K\|_1^2$. On the other hand,

$$\mathbb{E}_f \xi_h^2(\cdot)\mathbb{I}(\bar{\mathcal{A}}) \le \{\mathbb{E}_f [\xi_h(\cdot)]^{4s/(s+2)}\}^{(s+2)/(2s)} \{\mathbb{P}(\bar{\mathcal{A}})\}^{(s-2)/(2s)}$$

and, therefore,

$$\begin{aligned} &\|\mathbb{E}_f \xi_h^2(\cdot)\mathbb{I}(\bar{\mathcal{A}})\|_{2s/(s+2)} \\ &\quad \le \{\mathbb{E}_f (\|\xi_h\|_{4s/(s+2)})^{4s/(s+2)}\}^{(s+2)/(2s)} \{\mathbb{P}(\bar{\mathcal{A}})\}^{(s-2)/(2s)}. \end{aligned} \tag{54}$$

We derive from Theorem 1 in GL (2011) that there exists c_5 such that

$$\mathbb{E}_f (\|\xi_h\|_{4s/(s+2)})^{4s/(s+2)} \le c_5 (nV_h)^{-2s/(s+2)}. \tag{55}$$

Putting $b = x\|K\|_2$, $x \ge 2$, we have in view of (50)

$$\{\mathbb{P}(\bar{\mathcal{A}})\}^{(s-2)/(2s)} \le e^{c_2(1-x)(s-2)/(2s)}.$$

It leads, together with (54) and (55), to the following estimate:

$$\|\mathbb{E}_f \xi_h^2(\cdot)\mathbb{I}(\bar{\mathcal{A}})\|_{2s/(s+2)} \le c_6 (nV_h)^{-1} e^{c_2(1-x)(s-2)/(2s)}. \tag{56}$$

Finally, we obtain from (45), (52), (53) and (56)

$$\mathbb{E}_f \|\xi_h\|_s \ge (x\|K\|_2)^{-1}(nV_h)^{-1/2}\big[c_4 \mu \nu^{(2-s)/(2s)} - c_6 e^{c_2(1-x)(s-2)/(2s)}\big].$$

It remains to choose x sufficiently large and we come to the assertion of the theorem in the case $s > 2$.

4.4. *Proof of Theorem 4.* Let $f \in N_{s,d}(\alpha, L)$. It is easily checked [see, e.g., Proposition 3 in Kerkyacharian, Lepski and Picard (2001)] that bias of the estimator $\hat{f}_h$ is bounded as follows:

$$\|B_h(f,\cdot)\|_s \le C_1(d,s,l)L \sum_{j=1}^{d} h_j^{\alpha_j}.$$

Moreover, $\{\mathbb{E}_f \|\xi_h\|_s^q\}^{1/q} \le C_2 (nV_h)^{-\gamma_s}$. If we set the "oracle bandwidth" $h^* := (h_1^*, \dots, h_d^*)$ so that

$$[h_j^*]^{\alpha_j} := \left[\frac{C_2}{C_1}\right]^{\bar{\alpha}/(\gamma_s+\bar{\alpha})} L^{-\bar{\alpha}/(\gamma_s+\bar{\alpha})} n^{-\gamma_s\bar{\alpha}/(\gamma_s+\bar{\alpha})}, \qquad j = 1, \dots, d,$$

then $h^* \in \mathcal{H}$ and $\hat{f}_{h^*} \in \mathcal{F}(\mathcal{H})$ for large enough n. Hence, for any $f \in N_{s,d}(\alpha, L)$ we have that $\mathcal{R}_s[\hat{f}_{h^*}; f] \le C_3 \varphi_{n,s}(\bar{\alpha})$. Then we apply oracle inequalities of Theorems 1 and 2. Observe that by choice of constant $\varkappa_2$ in definition of $h^{\max}$ we guarantee that the remainder terms are negligibly small as $n \to \infty$ in comparison with the first terms in (10) and (11). This fact leads to the statement of the theorem.

4.5. *Proof of Theorem 5.* First we note that it suffices to prove the theorem only for $s \ge 2$. Indeed, since $\operatorname{supp}(f) \subseteq Q$, one has $\operatorname{supp}(\hat{f}_h) \subseteq Q'$ for any $\mathcal{X}_n$-measurable random vector $h \in \mathcal{H}$, where, in view of the assumptions imposed on the kernel K,

$$Q' = \{y \in \mathbb{R}^d : |y_i - x_i| \le 1/2, i = 1, \dots, d, x \in Q\}.$$

Here we have also used that $h^{\max} \in (0,1]^d$. Thus, for any density f and any $\mathcal{X}_n$-measurable random vector $h \in \mathcal{H}$

$$\operatorname{supp}(\hat{f}_h - f) \subseteq Q'$$

and, therefore, in view of Hölder inequality for any $s \in [1,2)$

$$\|\hat{f}_h - f\|_s \le [\operatorname{mes}\{Q'\}]^{(2-s)/(2s)} \|\hat{f}_h - f\|_2.$$

We conclude that for any $s \in [1,2)$ the estimation problem in the $\mathbb{L}_s$-norm can be reduced to the estimation problem in the $\mathbb{L}_2$-norm.

Let $f \in W_{p,d}(\alpha, L, Q)$ and $s \geq 2$. The standard computation (by the generalized Minkowski inequality and by the Hölder inequality along with the fact that f is compactly supported) yields the following bound on the $\mathbb{L}_s$-norm of the bias of $\hat{f}_h$:

$$\|B_h(f,\cdot)\|_s \leq C_1(d,s,l)L[\mathrm{mes}\{Q\}]^{(p-s)/(sp)} \sum_{j=1}^{d} h_j^{\alpha_j}.$$

Moreover, $\{\mathbb{E}_f \|\xi_h\|_s^q\}^{1/q} \leq C_2(nV_h)^{-1/2}$. If we set the "oracle bandwidth" $h^* := (h_1^*, \ldots, h_d^*)$ so that

$$[h_j^*]^{\alpha_j} := \left[\frac{C_2}{C_1}\right]^{2\bar{\alpha}/(2\bar{\alpha}+1)} \left(L[\mathrm{mes}\{Q\}]^{(p-s)/(sp)}\right)^{-2\bar{\alpha}/(1+2\bar{\alpha})} n^{-\bar{\alpha}/(2\bar{\alpha}+1)},$$

$$j = 1, \ldots, d,$$

then $h^* \in \mathcal{H}$ and $\hat{f}_{h^*} \in \mathcal{F}(\mathcal{H})$ for large enough n. Then the result follows by application of Theorems 1(ii) and 2.

APPENDIX

Proofs of Lemmas 1 and 2 follow directly from general uniform bounds on norms of empirical processes established in GL (2011). In our proofs below we use notation and terminology of this paper.

PROOF OF LEMMA 1. The statement is a direct consequence of Theorem 4 of Section 3.3 in GL (2011).

To apply this theorem one should verify Assumptions (W1), (W4) and (L) for the following classes of weights $\mathcal{W}^{(1)} = \{w = n^{-1}K_h : h \in \mathcal{H}\}$ and $\mathcal{W}^{(2)} = \{w = n^{-1}(K_h * K_\eta) : (h,\eta) \in \mathcal{H} \times \mathcal{H}\}$. The sets $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ are considered as images of $\mathcal{H}$ and $\mathcal{H} \times \mathcal{H}$ under transformations $h \mapsto n^{-1}K_h$ and $(h,\eta) \mapsto n^{-1}(K_h * K_\eta)$, respectively. The sets $\mathcal{H}$ and $\mathcal{H} \times \mathcal{H}$ are equipped with the distances

$$\mathrm{d}_1(h,h') = c_1 \max_{i=1,\ldots,d} \ln\left(\frac{h_i \vee h_i'}{h_i \wedge h_i'}\right),$$

$$\mathrm{d}_2[(h,h'),(\eta,\eta')] := c_2\{\mathrm{d}_1(h,h') \vee \mathrm{d}_1(\eta,\eta')\},$$

where c_1 and c_2 are appropriate constants depending on k_∞, L_K and d only [see formulas (9.1) and (9.2) in GL (2011)]. With this notation Lemma 9 of GL (2011) shows that Assumption (L) holds for both $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$. Moreover, Assumption (W1) holds trivially for both $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ with $\mu_* = V_{\max}$ and $\mu_* = 2^d V_{\max}$, respectively. Moreover, Assumption (W4) for both $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$ follows from formula (9.8) in GL (2011). Thus all conditions of Theorem 4 are fulfilled.

(i) We apply this theorem with $z = 1$ and $\varepsilon = 1$. We need to evaluate the constant $T_{3,\varepsilon}$ for $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$. If $N_{\mathcal{H},\mathrm{d}_1}(\varepsilon)$ denotes the minimal number of balls in the metric d_1 needed to cover $\mathcal{H}$, then formula (9.8) from GL (2011) shows that $N_{\mathcal{H},\mathrm{d}_1}(1/8) \le c_3 A_{\mathcal{H}}$, where c_3 depends on d only. Similarly, $N_{\mathcal{H}\times\mathcal{H},\mathrm{d}_2}(1/8) \le c_4 A_{\mathcal{H}}^2$. In addition, for

$$L_{\mathcal{H},\mathrm{d}_1}(\varepsilon) := \sum_{k=1}^{\infty} \exp\{2\ln N_{\mathcal{H},\mathrm{d}_1}(\varepsilon 2^{-k}) - (9/16)2^k k^{-2}\}$$

we have $L_{\mathcal{H},\mathrm{d}_1}(1) \le c_5 A_{\mathcal{H}}$. Similarly, $L_{\mathcal{H}\times\mathcal{H},\mathrm{d}_2}(1) \le c_6 A_{\mathcal{H}}^2$. Combining these bounds we come to the statement (i).

(ii) The second statement follows exactly in the same way from the above considerations. Theorem 4 of GL (2011) is again applied with $z = 1$ and $\varepsilon = 1$. □

PROOF OF LEMMA 2. The proof is by application of Theorem 7 from GL (2011). We need to calculate several quantities.

We start with the class $\mathcal{W}^{(1)}$. Here for $\vartheta_0^{(1)} = 10 D_s \mathrm{f}_\infty (L_K \sqrt{d})^{d/2}$ we have

$$\begin{aligned} C^*_{\xi,1}(y) &= 1 + 2\vartheta_0^{(1)}\{\sqrt{y}(V_{\max}^{1/s} + n^{-1/(2s)}) + y n^{-1/s}\} \\ &\le 1 + 2\vartheta_0^{(1)}\{2\sqrt{y} V_{\max}^{1/s} + y V_{\max}^{2/s}\}, \end{aligned}$$

where we have used that $V_{\max} \ge 1/\sqrt{n}$. If we set $y = \bar{y} := [4V_{\max}^{2/s}(\vartheta_0^{(1)} \vee 1)]^{-1}$, then $C^*_{\xi,1}(\bar{y}) \le 4$. We apply Theorem 7 with $\varepsilon = 1$ and $y = \bar{y}$. Condition $n V_{\min} > C_1 = [256 D_s^2]^{(s\wedge 4)/(s\wedge 4-2)}$ implies that

$$\bar{u}_1(\gamma) = 4\big[1 - 8 D_s (n V_{\min})^{1/(s\wedge 4)-1/2}\big]^{-1} \le 8.$$

Moreover, we note that condition $\bar{y} \le y_*^{(1)}$ follows from definition of $\bar{y}$ and $n \ge C_2$. In addition, $\tilde{T}_{1,\varepsilon}^{(1)} \le c A_{\mathcal{H}}^2 B_{\mathcal{H}}$. These facts imply (21) and (23).

The bounds (22) and (24) for $\mathcal{W}^{(2)}$ follow from similar computations. □

Acknowledgments. The authors thank two anonymous referees for useful comments and suggestions.

REFERENCES

BIRGÉ, L. (2008). Model selection for density estimation with $\mathbb{L}_2$-loss. Available at arXiv:0808.1416v2.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137. MR0523165

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation*: *The* L_1 *View*. Wiley, New York. MR0780746

DEVROYE, L. and LUGOSI, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *Ann. Statist.* **24** 2499–2512. MR1425963

DEVROYE, L. and LUGOSI, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.* **25** 2626–2637. MR1604428

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. Springer, New York. MR1843146

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974

GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. MR2543590

GOLDENSHLUGER, A. and LEPSKI, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. MR2449122

GOLDENSHLUGER, A. and LEPSKI, O. (2011). Uniform bounds for norms of sums of independent random functions. *Ann. Probab.* To appear. Available at arXiv:0904.1950v2.

HASMINSKII, R. and IBRAGIMOV, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18** 999–1010. MR1062695

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **98** 61–85. MR0591862

IBRAGIMOV, I. A. and KHAS'MINSKIĬ, R. Z. (1981). More on estimation of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **108** 72–88. MR0629401

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643. MR0238419

JOHNSON, W. B., SCHECHTMAN, G. and ZINN, J. (1985). Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *Ann. Probab.* **13** 234–253. MR0770640

JUDITSKY, A. and LAMBERT-LACROIX, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. MR2046772

KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. MR1863916

KERKYACHARIAN, G., PICARD, D. and TRIBOULEY, K. (1996). L^p adaptive density estimation. *Bernoulli* **2** 229–247. MR1416864

MASON, D. M. (2009). Risk bounds for kernel density estimators. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **363** 66–104. Available at http://www.pdmi.ras.ru/znsl/.

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

NIKOL'SKIĬ, S. M. (1969). *Priblizhenie Funktsii Mnogikh Peremennykh i Teoremy Vlozheniya*. Nauka, Moscow. MR0310616

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076. MR0143282

RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837. MR0079873

SAMAROV, A. and TSYBAKOV, A. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference* (V. Nair, ed.). *Ser. Biostat.* **3** 233–251. World Scientific, Hackensack, NJ. MR2416118

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London. MR0848134

DEPARTMENT OF STATISTICS
UNIVERSITY OF HAIFA
31905 HAIFA
ISRAEL
E-MAIL: goldensh@stat.haifa.ac.il

LABORATOIRE D'ANALYSE, TOPOLOGIE
ET PROBABILITÉS
UMR CNRS 6632
UNIVERSITÉ DE PROVENCE
39, RUE F. JOLIOT CURIE
13453 MARSEILLE
FRANCE
E-MAIL: lepski@cmi.univ-mrs.fr